Object Discovery in Videos as Foreground Motion Clustering

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Christopher Xie[1], Yu Xiang[2], Zaid Harchaoui[1], Dieter Fox[2,1]

[1]University of Washington, [2]NVIDIA

NVIDIA

PROBLEM

> Robots need the capability to discover unknown objects in arbitrary environments, e.g. new households or work spaces.

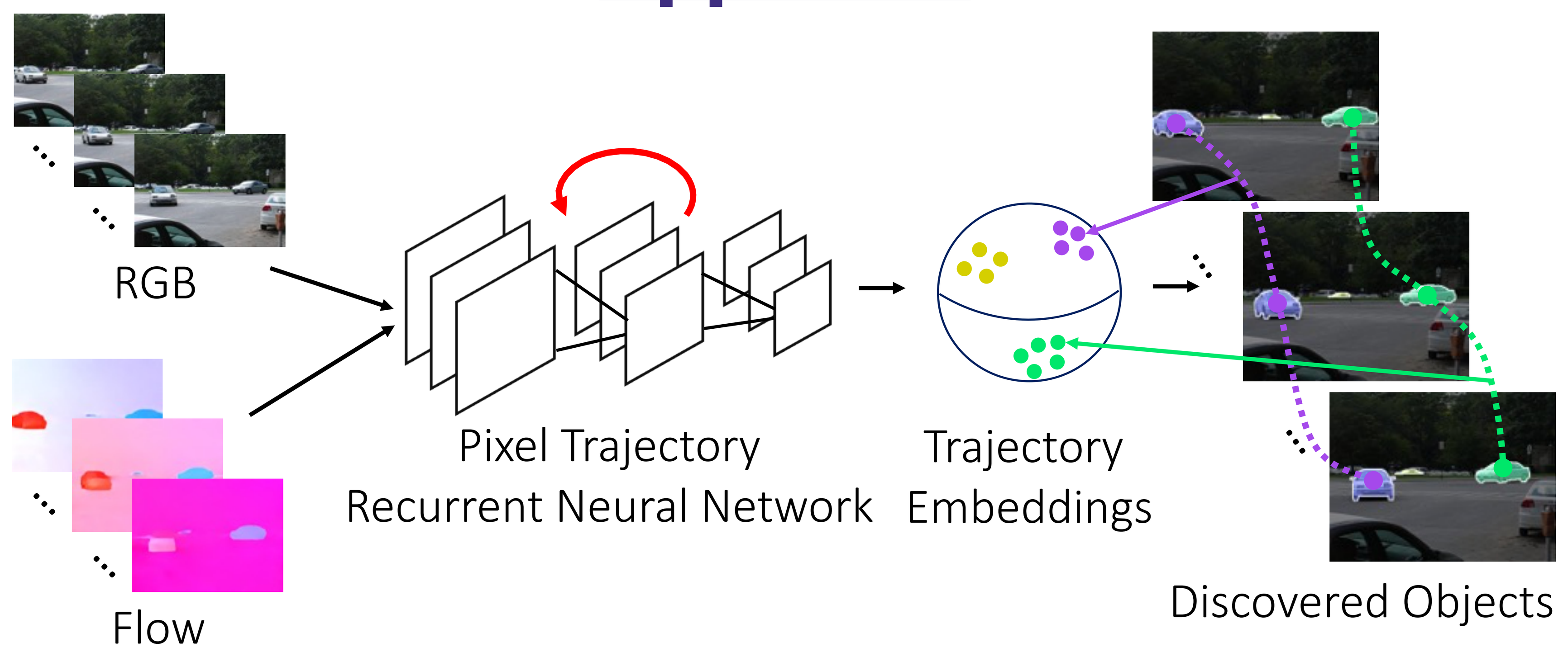
> Can a robot discover unknown objects by passively observing video?

Motivation

> We should be able to separate moving objects from background by looking at motion.

> Can we utilize these motion cues, along with appearance cues + pixel trajectories [1] (for temporal consistency) in an end-to-end framework to cluster video pixels into foreground objects?

Approach

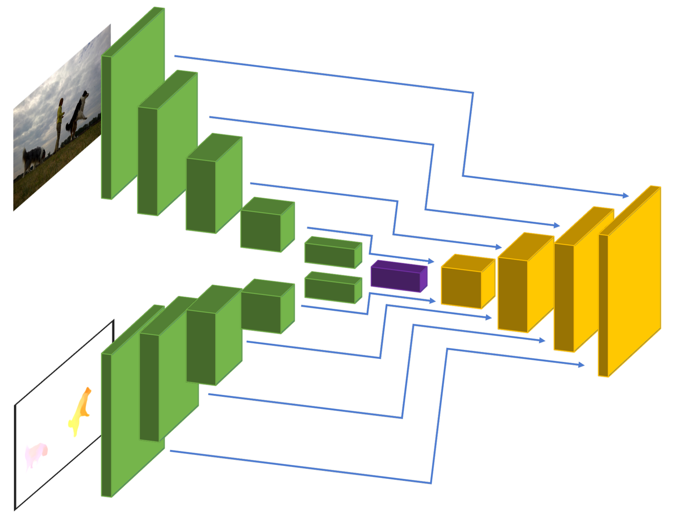


References

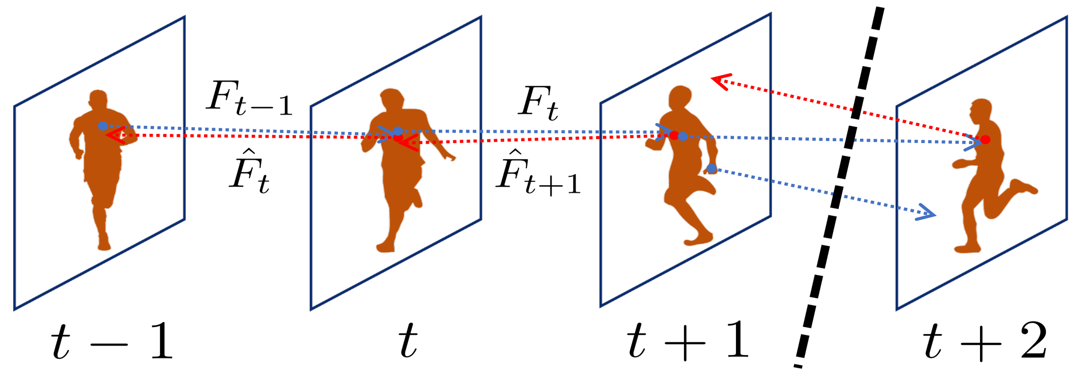
[1] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In European Conference on Computer Vision (ECCV), 2010.

METHOD

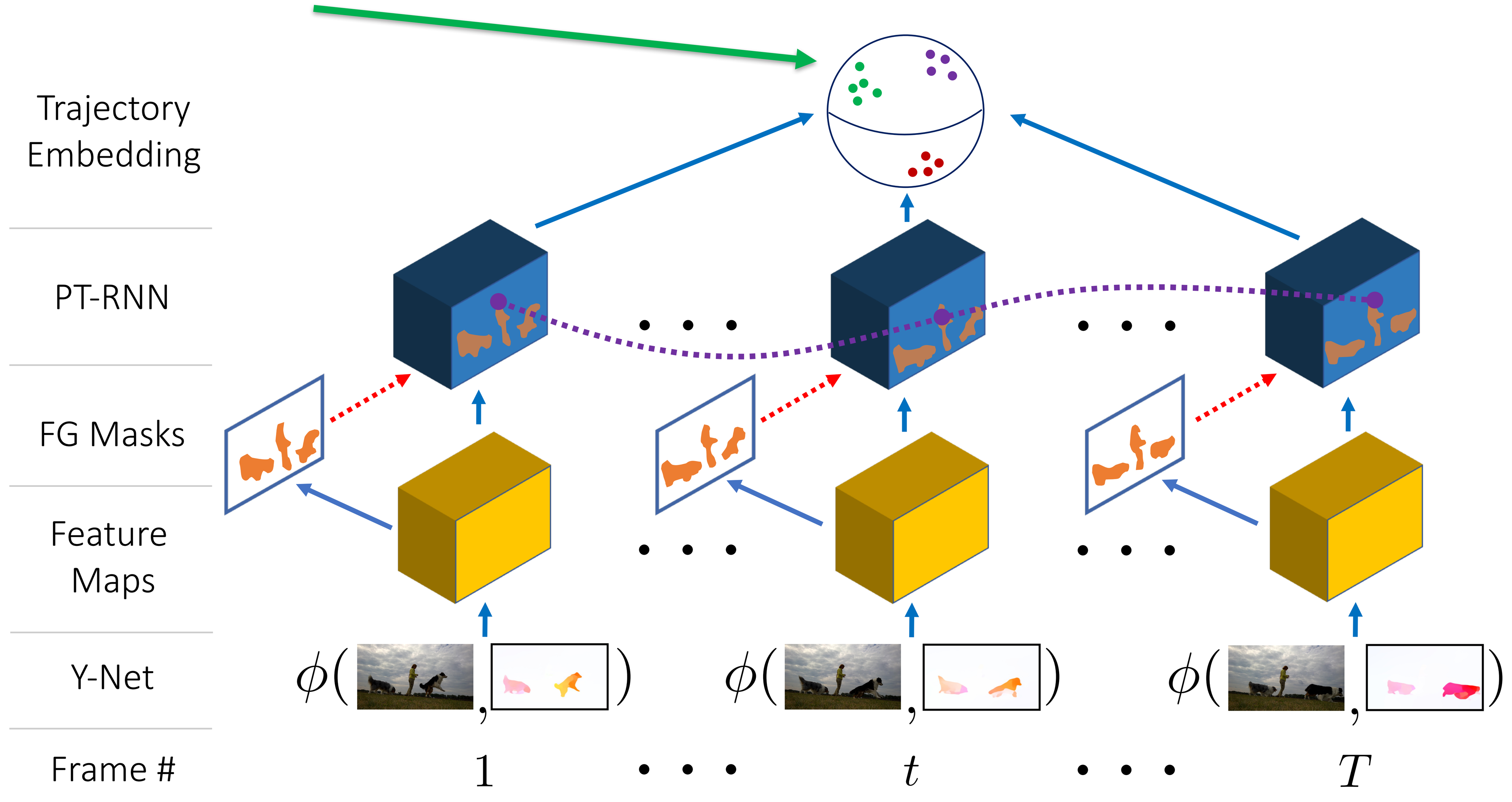
> Y-Net encoder-decoder to combine motion/appearance



> Foreground pixel trajectories



Pixel Trajectory RNN



EXPERIMENTS

Quantitative Results

> Motion segmentation benchmarks: Freiburg-Berkeley Motion Segmentation (FBMS), Complex Background (CB), Camouflaged Animal (CA)

		Video Foreground Segmentation							Multi-object Motion Segmentation			
		PCM [1]	FST [2]	NLC [3]	MPNet [4]	LVO [5]	CCG [6]	Ours	CVOS [7]	CUT [8]	CCG [6]	Ours
FBMS	P	79.9	83.9	86.2	87.3	92.4	85.5	90.3	72.7	74.6	74.2	75.9
	R	80.8	80.0	76.3	72.2	85.1	83.1	87.6	54.4	62.0	63.1	66.6
	F	77.3	79.6	77.3	74.8	87.0	81.9	87.7	56.3	63.6	65.0	67.3
	ΔObj	-	-	-	-	-	-	-	11.7	7.7	4.0	4.9
CB	P	84.3	87.6	79.9	86.8	74.6	87.7	83.1	60.8	67.6	64.9	57.7
	R	91.7	85.0	69.3	77.5	77.0	93.1	89.7	44.7	58.3	67.3	61.9
	F	86.6	80.6	73.7	78.2	70.5	90.1	83.5	45.8	60.3	65.6	58.3
	ΔObj	-	-	-	-	-	-	-	3.4	3.4	3.4	3.2
CA	P	81.9	73.3	82.3	77.8	77.6	80.4	78.5	84.7	77.8	83.8	77.2
	R	74.6	56.7	68.5	62.0	51.1	75.2	79.7	59.4	68.1	70.0	77.2
	F	76.3	60.4	72.5	64.8	50.8	76.0	77.1	61.5	70.0	72.2	75.3
	ΔObj	-	-	-	-	-	-	-	22.2	5.7	5.0	5.4
All	P	80.8	82.1	84.7	85.3	87.4	84.7	87.1	73.8	74.5	75.1	74.1
	R	80.7	75.8	73.9	70.7	77.2	82.7	86.2	54.3	62.8	65.0	68.2
	F	78.2	75.8	75.9	73.1	77.7	81.5	85.1	56.2	64.5	66.5	67.9
	ΔObj	-	-	-	-	-	-	-	12.9	6.8	4.1	4.8

[1] Bideau et al. ECCV 2016 [2] Papazoglou et al. ICCV 2013 [3] Faktor & Irani, BMVC 2014 [4] Tokmakov et al. CVPR 2017
[5] Tokmakov et al. ICCV 2017 [6] Bideau et al. CVPR 2018 [7] Taylor et al. CVPR 2015 [8] Keuper et al. ICCV 2015

> Architecture/Dataset ablation

	Multi-object				Foreground		
	P	R	F	ΔObj	P	R	F
conv PT-RNN	75.9	66.6	67.3	4.9	90.3	87.6	87.7
standard PT-RNN	72.2	66.6	66.0	4.27	88.1	89.3	87.5
convGRU PT-RNN	73.6	63.8	64.8	4.07	89.6	85.8	86.3
per-frame embedding	79.9	56.7	59.7	11.2	92.1	85.4	87.4
no FG mask	63.5	60.3	59.6	1.97	82.5	85.7	82.1
no SCM	70.4	65.5	63.2	3.70	89.3	89.1	88.1
no pre-FT3D	70.2	63.6	63.1	3.66	87.6	88.2	86.3
no DAVIS-m	66.9	63.6	62.1	2.07	87.1	86.9	85.2

Performance measured in IoU

	FT3D	DAVIS	FBMS
Y-Net	0.905	0.701	0.631
Early Fusion	0.883	0.636	0.568
Late Fusion	0.897	0.631	0.570

Qualitative Results